Characterizing the semantic composition of the UMLS Metathesaurus over time

Olivier Bodenreider, M.D., PhD and Lee Peters, M.S.
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
{olivier.bodenreider|lee.peters}@nih.gov

Motivation. The UMLS Metathesaurus has grown dramatically over the past fifteen years. From 2002 to 2015, the number of concepts has increased from about 777,000 to 3.1 million, a 4-fold increase. It is difficult to infer the semantic composition of the UMLS from the list of its sources. While some source vocabularies contribute concepts from a single semantic category (e.g., anatomical entities in the Foundational Model of Anatomy), others reflect a wide range of semantic categories (e.g., SNOMED CT). Moreover, the integration of a given term from a source vocabulary does not always result in new concepts, since this term may end up as a synonym for an existing concept. In this investigation, we leverage the semantic groups to characterize the semantic composition of the UMLS Metathesaurus and its evolution over time.

Methods. Each Metathesaurus concept is assigned at least one of the 127 semantic types. Semantic types are grouped into fifteen semantic groups, which represent broad subdomains of biomedicine, such as *Anatomy*, *Chemicals and Drugs*, and *Disorders*. The UMLS semantic groups have been used to create semantic profiles for source vocabularies, but can also be applied to the Metathesaurus as a whole. For each edition of the UMLS (2002-2015), we compute the distribution of the Metathesaurus concepts with respect to the 15 semantic groups.

Findings. As shown in Figure 1, the proportion of a few semantic groups has changed markedly between 2002 and 2015. ***Chemicals & Drugs***. The composition of the early Metathesaurus versions was heavily dominated by chemical concepts, especially from MeSH. Although the number of chemical concepts has almost doubled during this period, it has grown at a slower pace than that of other groups. ***Living Beings***. This group mostly represents organisms (mainly from the NCBI taxonomy) and has grown from 30,000 to nearly 1M concepts. It is the fastest growing group and now represents 30% of all Metathesaurus concepts. ***Disorders and Procedures***. The integration of a single large and fine-grained vocabulary can be responsible for major shifts in composition. The growth of disorder concepts between 2009 and 2011 is attributable to the integration of two large and fine-grained vocabularies, MEDCIN (2009) and ICD10-CM (2011). This is also the case for procedure concepts in 2009, when ICD10PCS was added. In contrast, the integration of SNOMED CT in 2003-2004 is silent, because, although extensive and detailed, its content was already largely represented by SNOMED International and the Read Codes.

For information about the UMLS semantic groups, see: https://semanticnetwork.nlm.nih.gov/.

Acknowledgments: This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

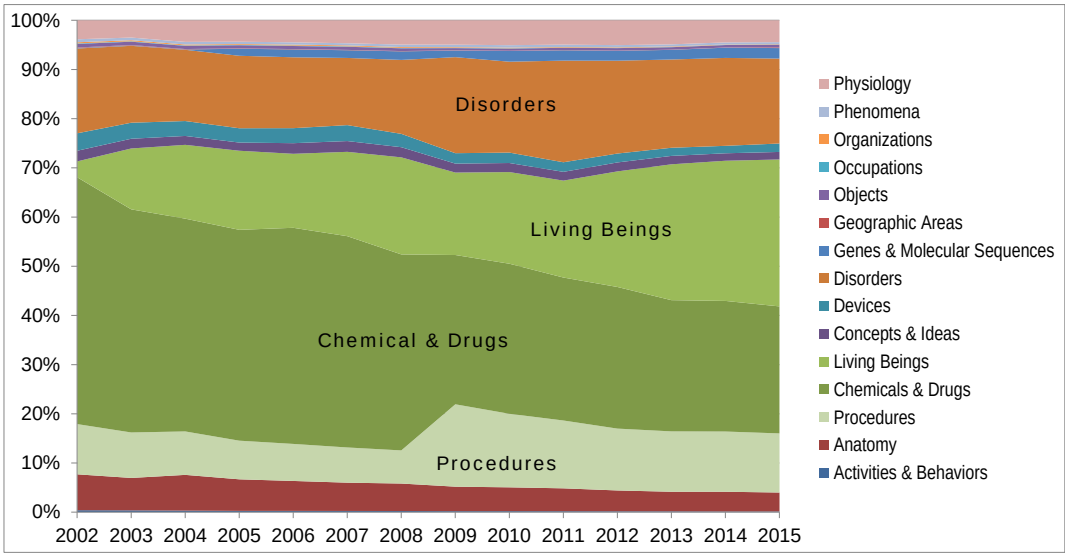

Figure 1. Evolution of the semantic group distribution in the UMLS Metathesaurus